Principled Evaluation of Differentially Private Algorithms

Gerome Miklau (UMass Amherst)

joint work with

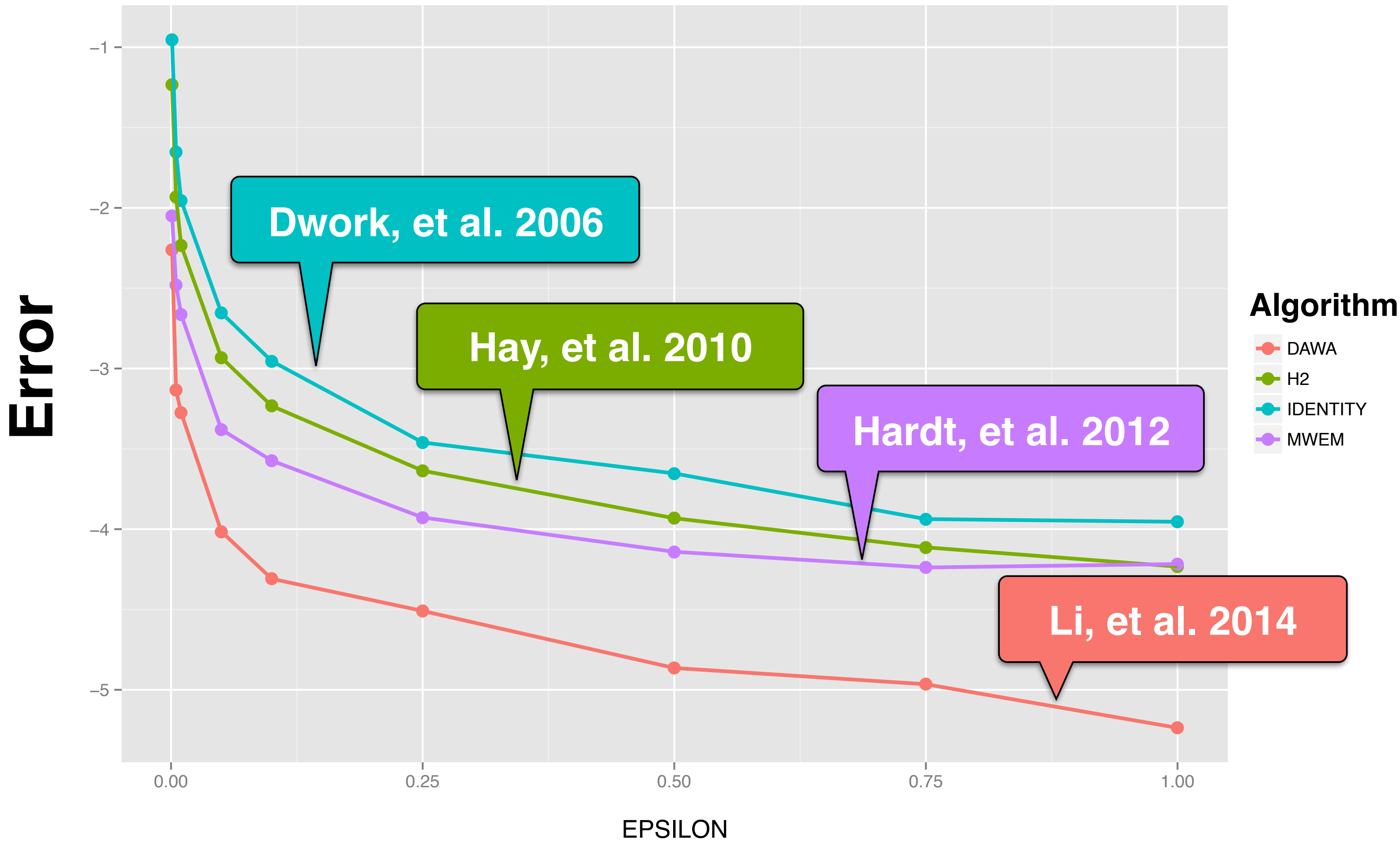
Ashwin Machanavajjhala (Duke)

Michael Hay (Colgate)

Dan Zhang (UMass Amherst)

Yan Chen (Duke)

Impressive progress (?)

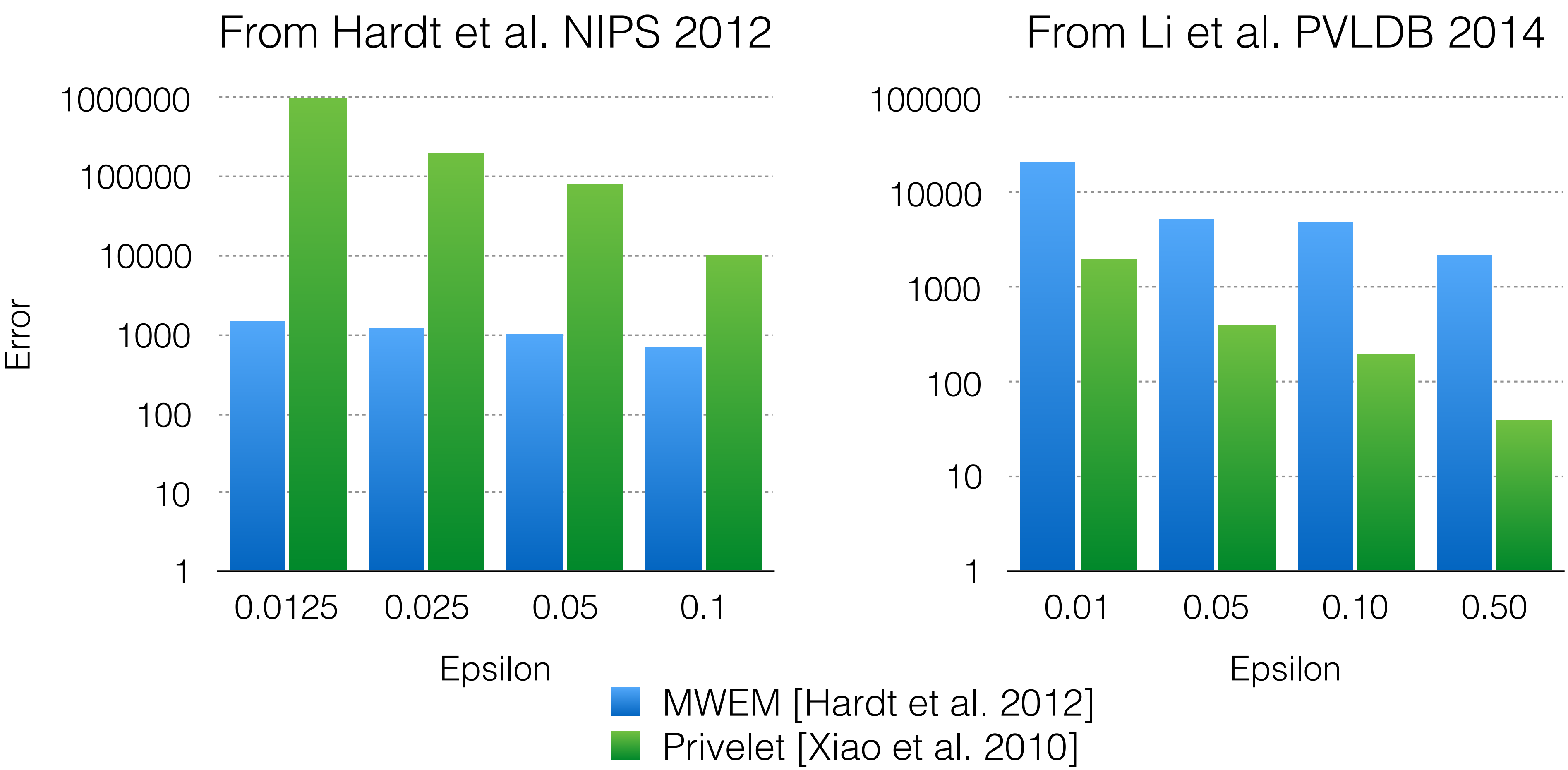


Task: 2000 range queries; Dataset: trace;
Scale = 10000; domain size = 4096

Obstacles to adoption

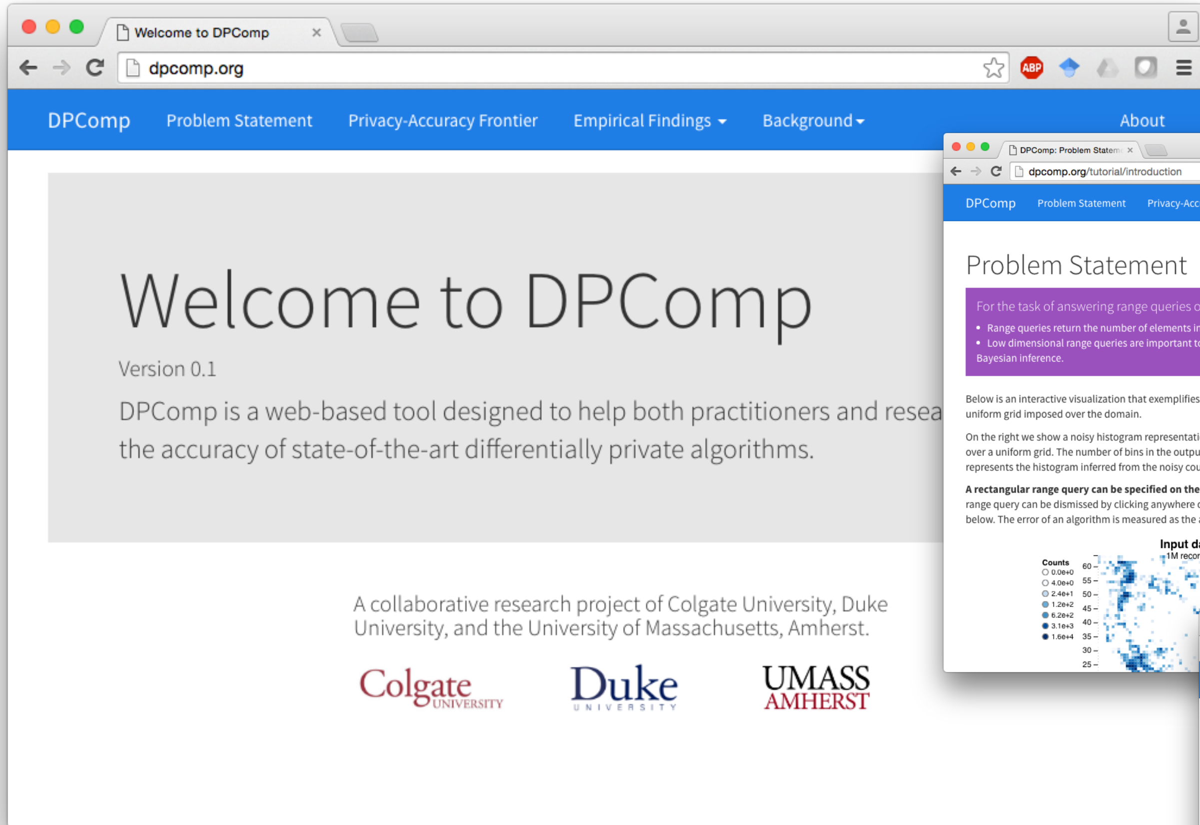
- Privacy researchers have adopted an idealized and simplistic view of a data analyst's workflow, often ignoring:
  - data representation, data cleaning, model selection, feature selection, algorithm tuning, iterative analysis.
- Practical performance of privacy algorithms is **opaque to users** and, in some cases, **poorly understood by researchers**.
  - The best algorithm for a task may depend on: setting of epsilon, "amount" of data, tunable algorithm parameters, data pre-processing (cleaning, representation)
  - Algorithm performance can be **data-dependent** because algorithms adapt or introduce bias.
- The **research community lacks rigorous methodology** for empirical evaluation.

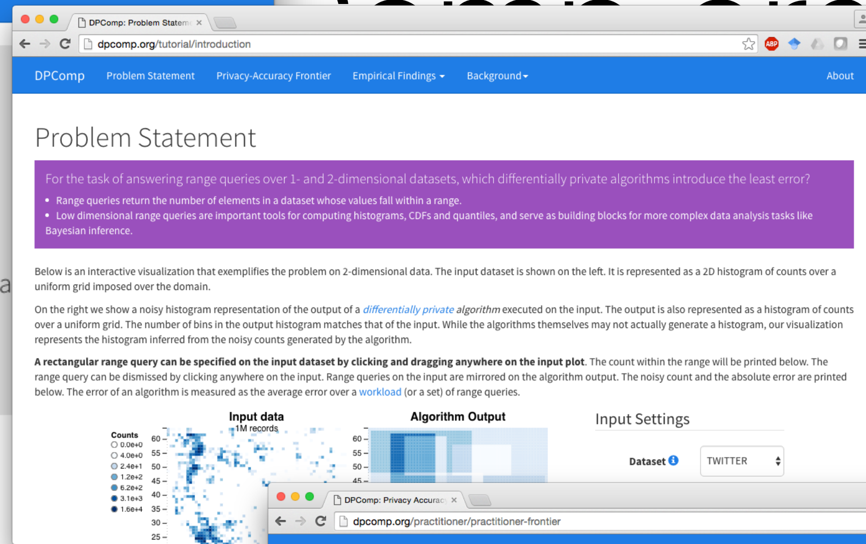
Conflicting results

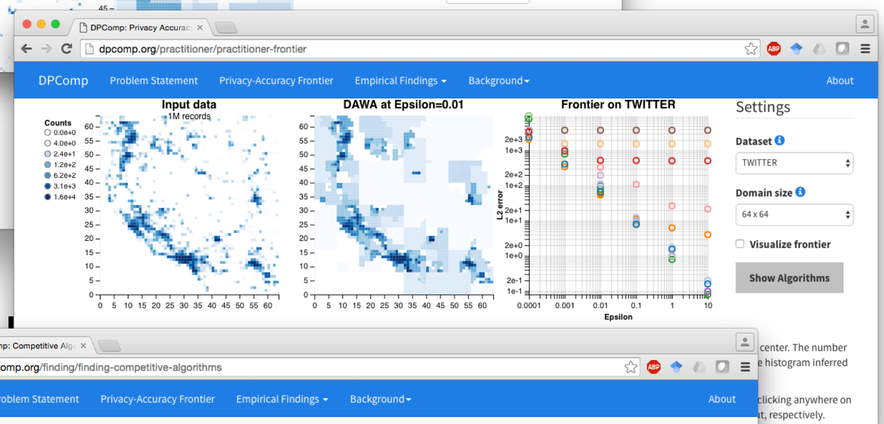


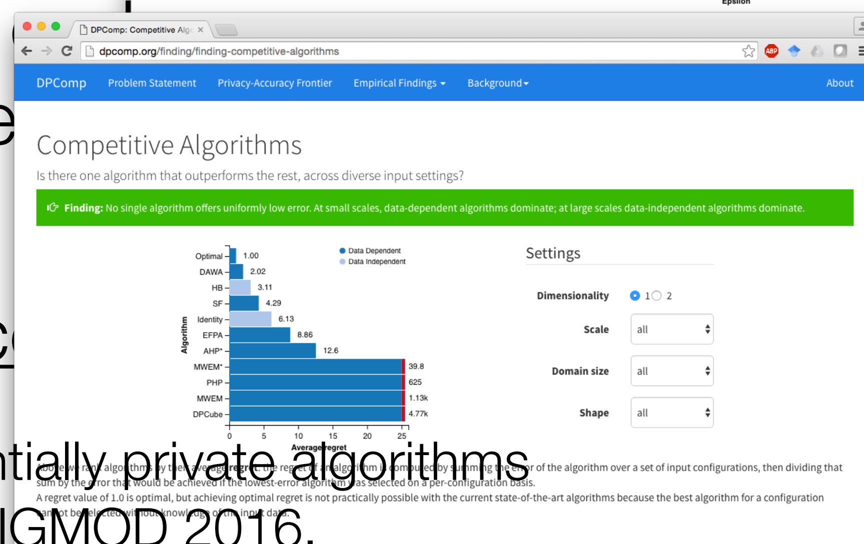
Our inspiration

- **Self-critique in machine learning:**
  - E.g. simple classifiers work well in practice; algorithm improvements dwarfed by ignored real-world factors; extreme focus on UCI datasets. Holte 1993, Hand 2006, Carbonell 1992, Wagstaff 2012.
- **Value of benchmarks:**
  - “When a field has good benchmarks, we settle debates and the field makes rapid progress.” David Patterson, CACM 2012.
- **MLcomp:**
  - Automated help for practitioners selecting algorithms for ML tasks.









workloads of range queries:

- 15 published algorithms
  ~8,000 distinct experime

- A companion website: dpc

Principled evaluation of differentially private algorithms using DPBench. In SIGMOD 2016.

Remainder of the talk

- 10 Principles
- Setup for benchmark study
- Overview of findings
- Open problems
- Our ongoing research efforts (motivated by dpcomp)

Evaluation principles

Diversity of inputs (Principles 1-4)

Diverse epsilon, diverse input data (scale, shape, domain size)

End-to-end privacy (Principles 5-7)

private pre- and post-processing; no free parameters; no side information.

Sound evaluation of output (Principles 8-10)

measure error variability; measure bias; compare algorithms using inputs that result in reasonable privacy and accuracy.

Task: Answering range queries

Sensitive Dataset

name	gender	nationality	grade
Alice	Female	US	91
Bob	Male	Canada	84
Carlos	Male	Peru	82
Darmesh	Male	India	97
Eloise	Female	France	88
Faith	Female	US	78
G			

Attributes
(dimensions)

{gender, grade}

Workload of Range Queries

"Number of A female students"
(count where gender=female and grade >= 93)

"Number of C students"
(count where gender=* and 70 <= grade < 80)

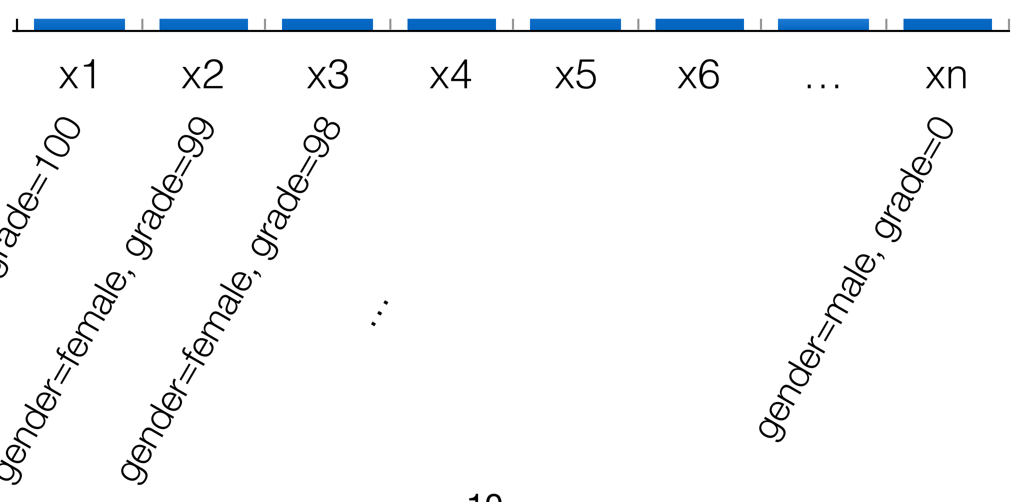
…

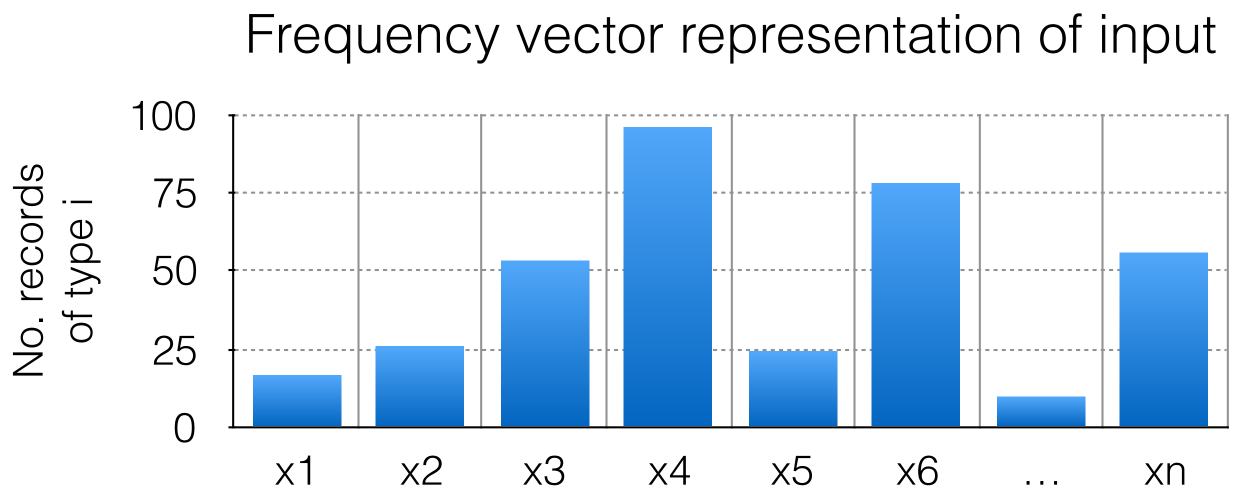
Task: Given workload of counting range queries on 1-2 dimensions, compute answers under ε-differential privacy

Diverse datasets

Principle: Data-dependent algorithms should be evaluated on a *diverse* set of inputs

Frequency vector representation of input





Properties:

- **domain size**: length of frequency vector
- **scale**: total number of records in database
- **shape**: the frequency vector normalized by scale.

Desideratum: datasets that are diverse with respect to all three properties.

Data generation

Systematically control for domain size and scale

Shape

Collect many real-world datasets

Dataset name	Original Scale	% Zero Counts	Previous works
1D datasets			
ADULT	32,558	97.80%	[10, 15]
HEPPH	347,414	21.17%	[15]
INCOME	20,787,122	44.97%	[15]
MEDCOST	9,415	74.80%	[15]
TRACE	25,714	96.61%	[1, 11, 27, 29]
PATENT	27,948,226	6.20%	[15]
SEARCH	335,889	51.03%	[1, 11, 27, 29]
BIDS-FJ	1,901,799	0%	new
BIDS-FM	2,126,344	0%	new
BIDS-ALL	7,655,502	0%	new
MD-SAL	135,727	83.12%	new
MD-SAL-FA	100,534	83.17%	new
LC-REQ-F1	3,737,472	61.57%	new
LC-REQ-F2	198,045	67.69%	new
LC-REQ-ALL	3,999,425	[illegible]	[illegible]
LC-DTIR-F1	[illegible]	[illegible]	[illegible]
LC-DTIR-F2	[illegible]	[illegible]	[illegible]
LC-DTIR-[illegible]	[illegible]	[illegible]	[illegible]
2D datasets			
BJ-CABS-S	[illegible]	[illegible]	[12]
BJ-CABS-E	[illegible]	[illegible]	[12]
GOWALLA	[illegible]	88.92%	[21]
ADULT-2D	32561	99.30%	[10]
SF-CABS-S	464040	95.04%	[20]
SF-CABS-E	464040	97.31%	[20]
MD-SAL-2D	70526	97.89%	new
LC-2D	550559	92.66%	new
STROKE	19435	79.02%	new

Domain size

Coarsen domain

dom(grades)

[0,100]

or

{A,B,C,D,F}

or

Scale

Sample with replacement

Input: real dataset *D*, domain *dom*, target scale *m*

requency vector **x'**

bucket

x' → Empirical dist. **p**

normalize

p, m → Frequency vector **x** of scale *m*

sample

See paper for details

Measuring error

Definition 7 (Scaled Average Per-Query Error). *Let* $\mathbf{W}$ *be a workload of* q *queries,* $\mathbf{x}$ *a data vector and* $s = \|\mathbf{x}\|_1$ *its scale. Let* $\hat{\mathbf{y}} = \mathcal{K}(\mathbf{x}, \mathbf{W}, \epsilon)$ *denote the noisy output of algorithm* $\mathcal{K}$*. Given a loss function* L*, we define* scale average per-query error *as* $\frac{1}{s \cdot q} L(\hat{\mathbf{y}}, \mathbf{W}\mathbf{x})$.

Example (scaled error):

	Scale	**Absolute Error**	**Scaled Absolute Error**
Dataset 1	1,000	100	0.100
Dataset 2	100,000	100	0.001

Scaled error is also error in units of a "population percentage"

Algorithms considered

Algorithm
Data-independent
IDENTITY [7]
PRIVELET [25]
H [11]
H_b [22]
GREEDY H [15]
Data-dependent
UNIFORM
MWEM [10]
MWEM*
AHP [29]
AHP*
DPCUBE [26]
DAWA [15]
QUADTREE [4]
UGRID [21]
AGRID [21]
PHP [1]
EFPA [1]
SF [27]

Data Independent

Data Independent

Laplace mechanism on frequency vector

Extensions of Laplace mechanism

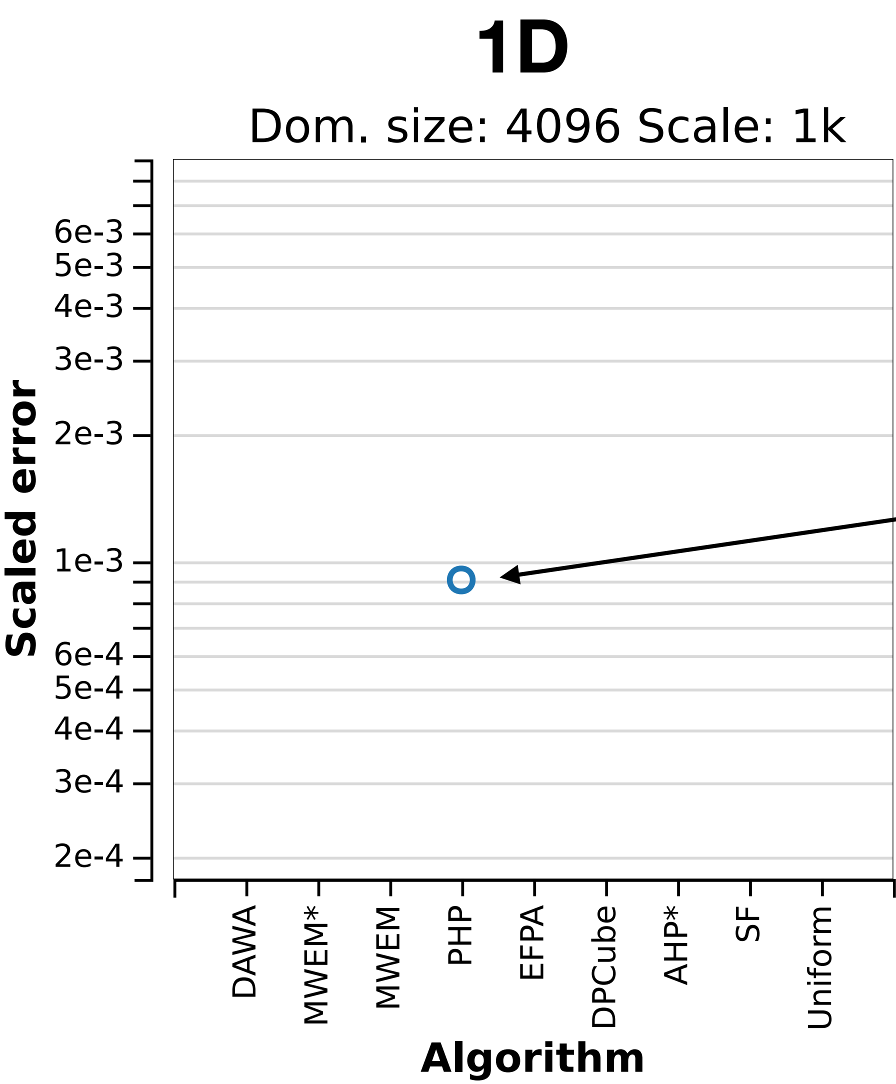
Noisy total count; assume uniformity.

Private partitioning; measurement over reduced domain.

2D-grid based techniques

Findings

Variation with shape



Error for a dataset

Dimensions: 1

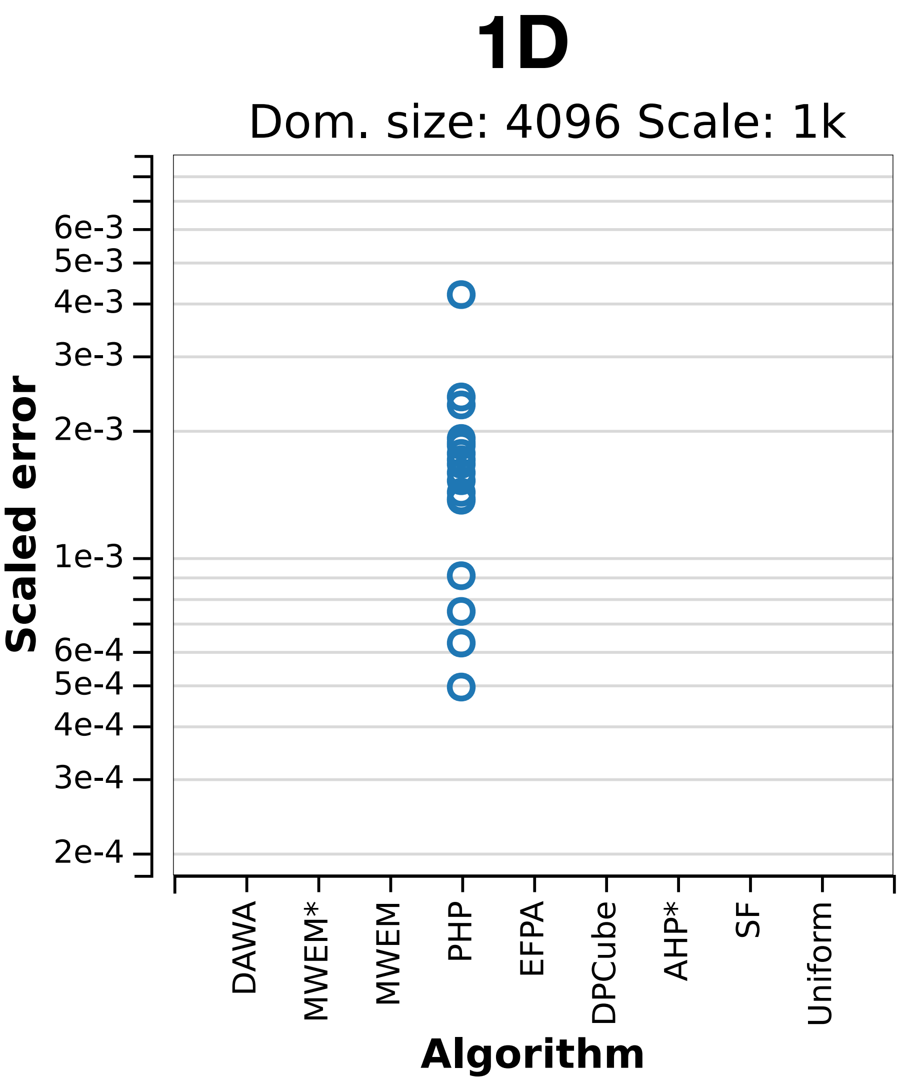
Shape: Patent

Domain size: 4096

Scale: 1000

(ε=0.1 throughout)

Variation with shape

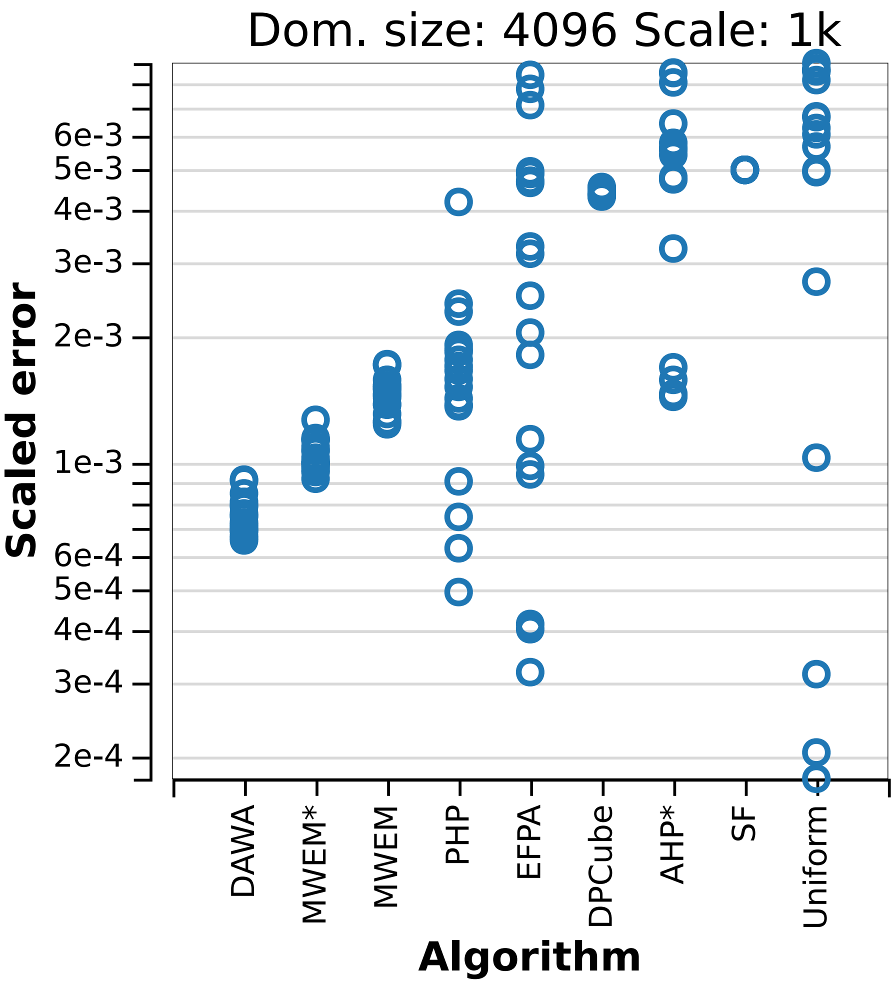


Variation across shape

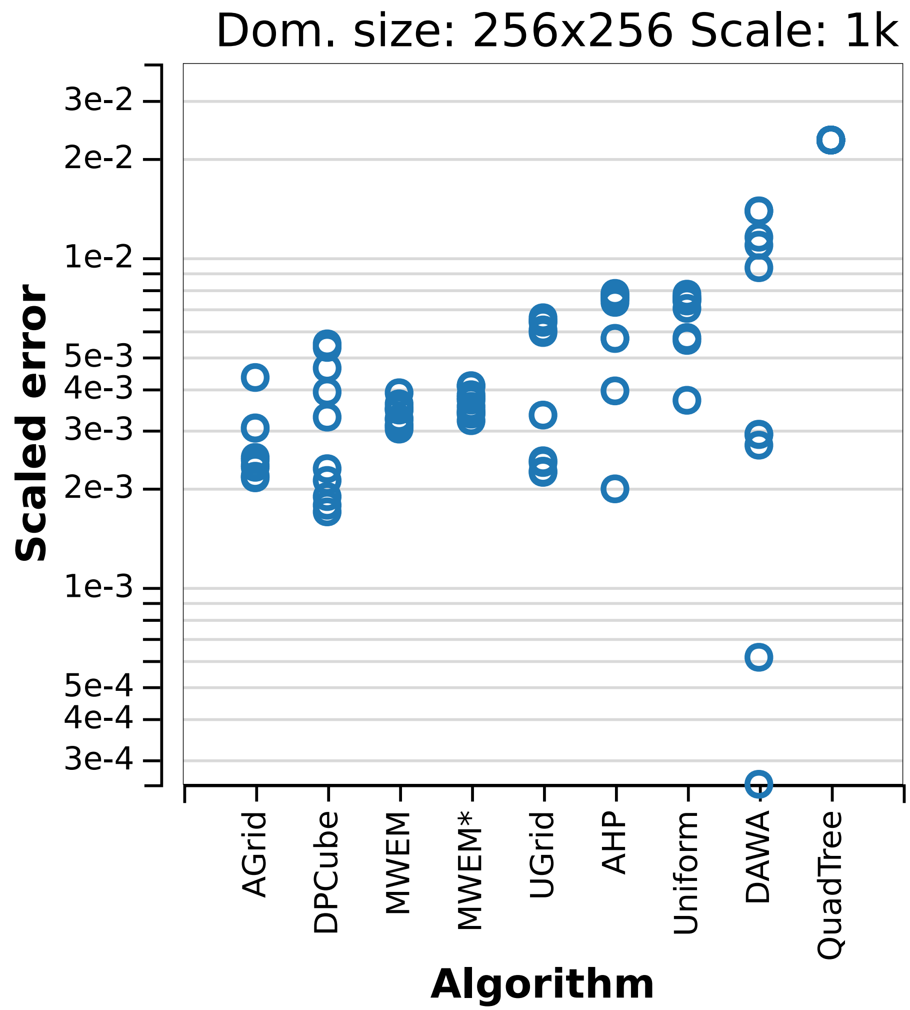
(for fixed dimension, domain size, scale)

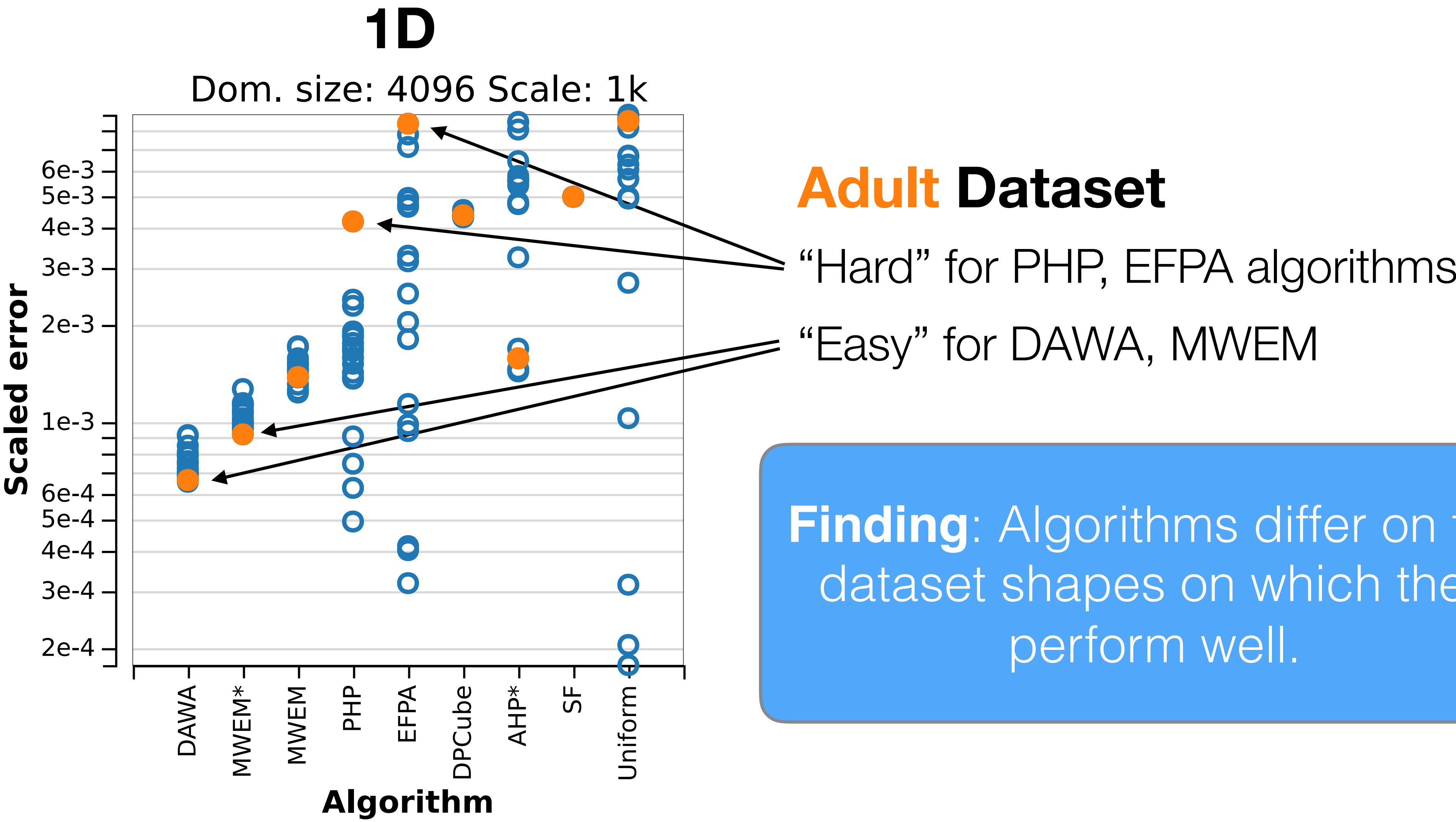
Finding: Algorithm error varies significantly with dataset shape

1D



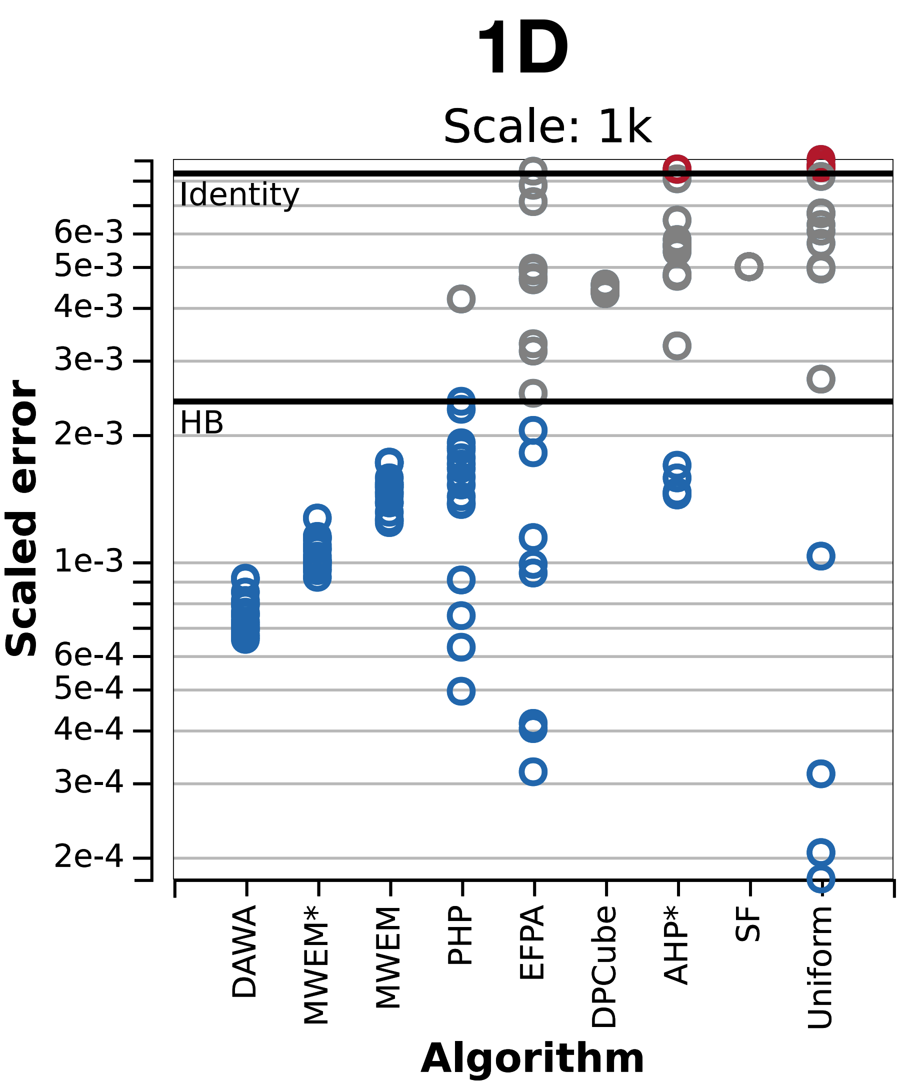
2D





Finding: Algorithms differ on the dataset shapes on which they perform well.

Data-independent alternatives



Data independent yardsticks

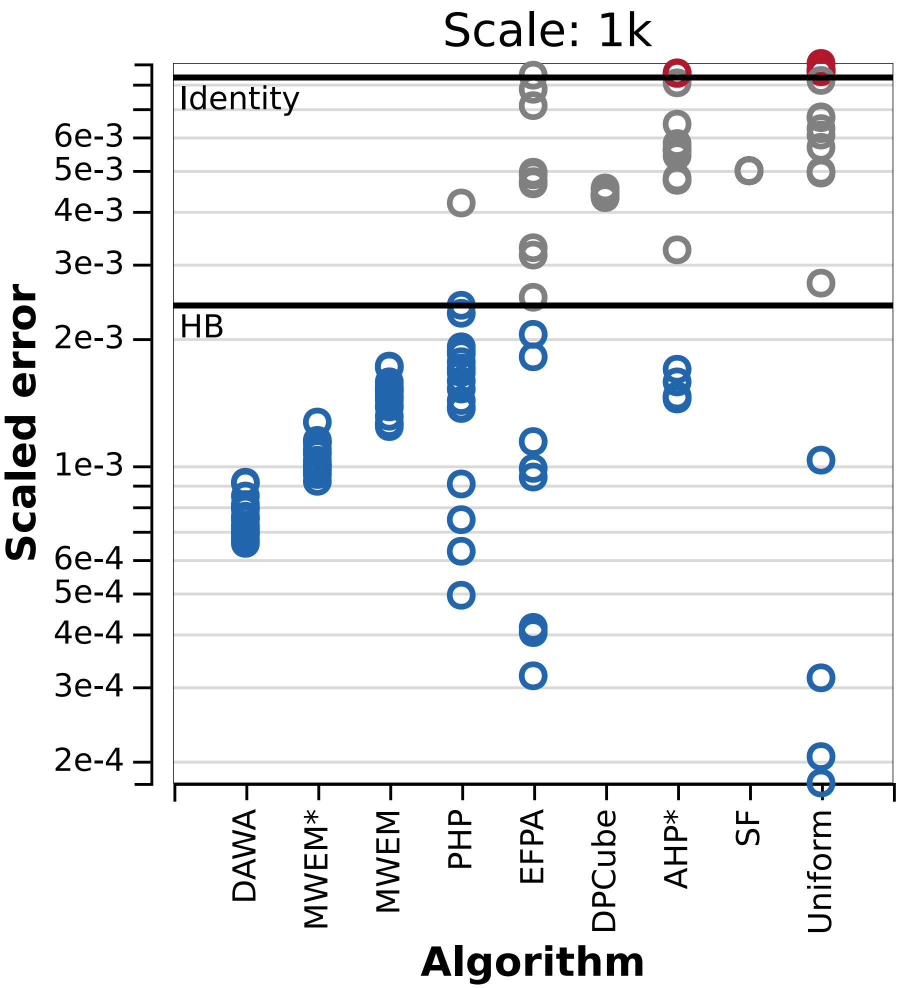
Identity: Laplace noise added to frequency vector **x**

HB: hierarchy of noisy counts

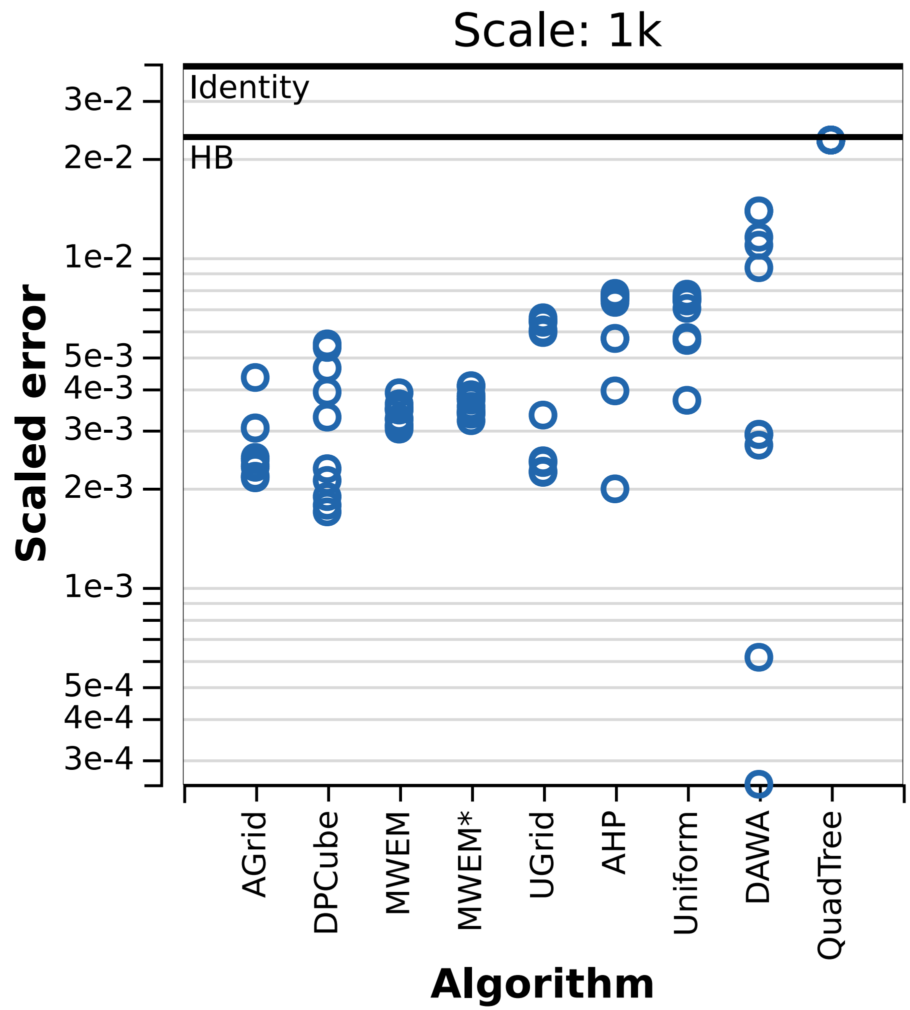
[Qardaji et al. ICDE 2013]

Finding: Data-dependence can offer significant improvements in error (at smaller scales or lower epsilon).

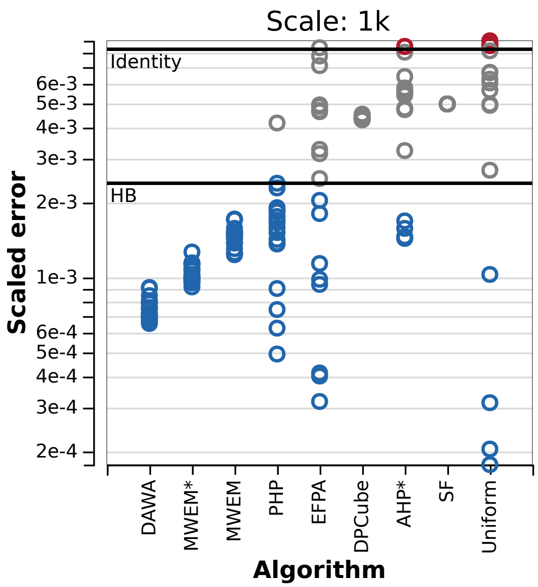
1D



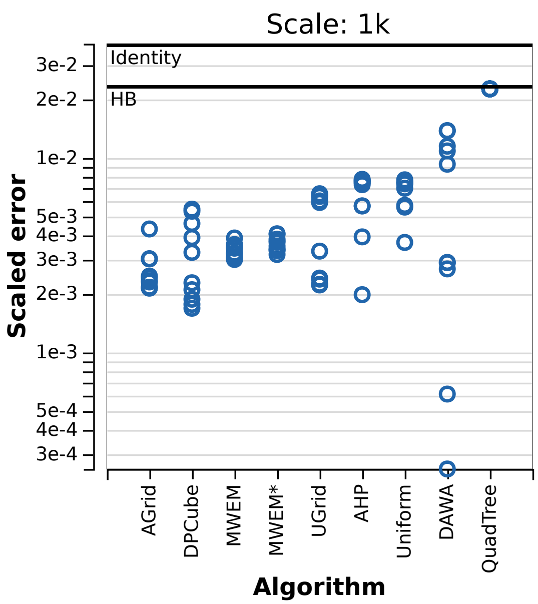
2D



1D



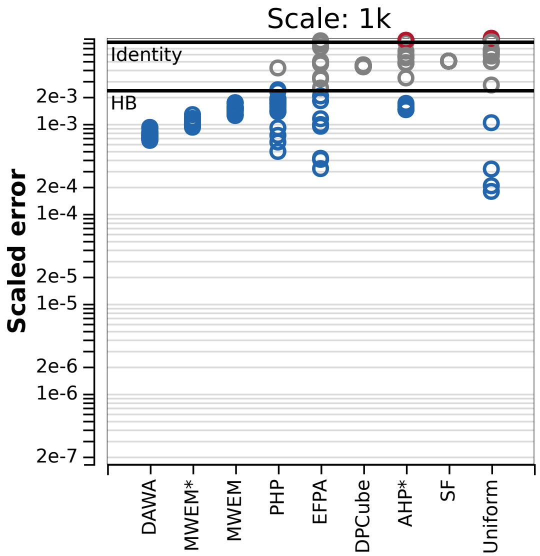
2D



Increasing scale →

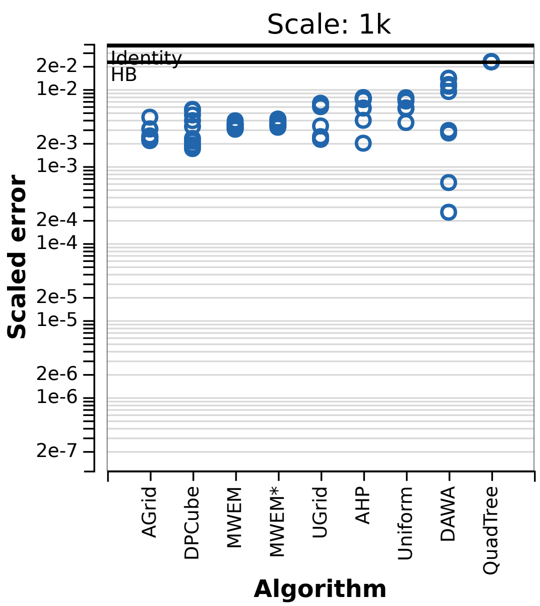
Finding: Some data-dependent algorithms fail to offer benefits at larger scales (or higher epsilons).

1D



Increasing scale →

2D



Review of Findings

- **No best algorithm:**
  - No single algorithm offers uniformly low error.
- **Significant variation with shape**
  - Algorithm error varies significantly with dataset shape and algorithms differ on the dataset shapes on which they perform well.
- **Significant trade-offs with “signal strength”**
  - Data-dependence can offer significant improvements in error, at smaller scales or lower epsilon values, but some data-dependent algorithms fail to offer benefits at larger scales or higher epsilons.
- **Failure to beat baselines**
  - Many algorithms are beaten by the IDENTITY baseline at large scales, in both 1D and 2D. At low scales, many algorithms result in error rates that are comparable to, or worse than, the Uniform baseline.

A few open questions

- Robust and private algorithm selection
  - See: Chaudhuri & Vinterbo, NIPS 2013, and our recent work “Pythia” SIGMOD 2017.
- Specialized data-dependent algorithms, or universal algorithms that can exploit structure in data?
- Error bounds for data-dependent algorithms
- Theory for non-worst case and for realistic parameters (concrete vs. asymptotic analysis)
- Richer, more complete benchmarks?